Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 11:14:07
REFERENCE NO: 228

Author Names & Affiliations

- Stephen Cross - Georgia Institute of Technology

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

Various, as submission is on behalf of the faculty and staff of Georgia Tech

Title of Submission

From Human Capital, Security and Standardization: Investments in Cyberinfrastructure Impact 21st Century Science and Discovery

Abstract (maximum ~200 words).

As a major technological research institution, the Georgia Institute of Technology has direct experience with cyberinfrastructure at all levels. Investment in heterogeneous, sustainable, scalable, secure, and compliant cyberinfrastructure is critical to enable future discoveries. Significant resources are needed to address the storage, network bandwidth, and massive computational power required for simulation and modeling across multiple scales. Data-centric computing is also vital, necessitating high-throughput analysis and mining of massive datasets, as well as the ongoing demand for low cost, long-term, reliable storage. Sustained investment in cybersecurity will support sharing of datasets along with greater multi-institution and multi-disciplinary research collaboration.
A significant investment in software engineering will enable researchers to leverage the promise offered by public-private, multi-cloud based cyberinfrastructure and emerging new architectures.
Some of the greatest risks are an inability to meet workforce demand and the lack of a sustainable funding model. Addressing these issues includes maximizing the steady pipeline of students entering science and engineering careers; creating professional retooling programs; building specialized local and regional teams; and leveraging a range of investment sources including federal, state, municipal and local entities, as well as public-private partnerships (e.g. academic and industry, government and corporate).

Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

Science and engineering research is the key to understanding everything in our universe and the best way we can improve the human condition. We are on the cusp of answering fundamental questions in the physical sciences, life sciences, social sciences, and mathematical and computational sciences. As our understanding deepens, we can leverage our basic fundamental knowledge to develop innovative and creative technologies that help drive solutions to the most pressing global problems ¬– all enabled by advances in

cyberinfrastructure.

As a major technological research institution, the Georgia Institute of Technology, which includes academic units and the Georgia Tech Research Institute (GTRI), has direct experience with many of the current and emerging research challenges facing today's scientists and engineers. To present a balanced response, we conducted a survey of more than 100 researchers, soliciting a broad cross section of faculty who are very familiar with existing cyberinfrastructure, as well as other researchers and staff who are working on technologies that provide a glimpse into the future. We received more than 50 responses, an indication of broad engagement in future cyberinfrastructure investments across campus. Below is an overview of four of the many research challenges that will both define and utilize current and future cyberinfrastructure over the coming decade and beyond.

CHALLENGE 1: Leveraging 21st century data to discover 21st century science

We now have the ability to collect and analyze several orders of magnitude more data than has ever been possible. But, there are inherent challenges in harnessing and making good use of that data across all areas of science and engineering. Compounded by the complexity and diversity of scientific data, these challenges create a bottleneck to future discovery that must be overcome.

Twenty-first century data is much broader than the traditional large instrument datasets. Large instruments, such as NSF's LASER Interferometric Gravitational Wave Observatory, acquire data using a traditional model where the analysis occurs after collection. Significant resources for storage, network bandwidth, and massive computational power are needed to analyze and mine such datasets.

Beyond large instruments, an ever-increasing quantity of networked devices provide a very different type of widely distributed and loosely correlated data streams, generating a nearly unmanageable "data deluge" generated by sensor networks. This raw observational data is of high scientific value, but we are often not able to leverage it due to lack of widely distributed storage, transport layers, and computational capabilities that are commensurate with the widely distributed sources of data.

Data is being generated across all disciplines. From just a subset of our survey, we found needs in processing and analysis of image and video data for medical applications, genomic data with sequencing technology, social science data from digital sources and multiple databases, and models and simulations that both generate data and need access to external data. Our ability to handle data well is clearly one of the greatest research challenges that cuts across all science and engineering.

CHALLENGE 2: Reducing friction at the interface between technology and humans

We routinely use technology that would be unimaginable to previous generations of scholars. New architectures, robotic systems and improvements in computing and storage, ranging from advanced chip design to new protocols, can help the working scientist, but only if he or she can easily use them. As the technology advances, scientists and engineers are not always able to take full advantage. One example is the demand for increased access to GPUs and other hybrid computing platforms, but insufficient resources to migrate codes.

Finding ways to take the human out of the loop regarding technical improvements will make it easier for technology advances to positively impact scientific research. There are several potential mechanisms to achieve this. We can develop abstraction layers, advanced machine learning tools, and artificial intelligence approaches, reducing the need for researchers to constantly refactor their codes in order to leverage changes in computational architecture.

New architectures will drive algorithmic improvements, speeding up core mathematical functions and developing robust uncertainty quantification. Computer science advances in deep learning show great potential as well. It will only be through improved software and advanced algorithms that we will be able to continue to scale codes to leverage these advanced architectures.

The seamless interaction between humans and robotic systems needs improvement, especially to facilitate discovery in inhospitable environments. Many researchers would benefit through better leveraging of "edge-computing" in terms of data collection and device control, as well as in-memory computing and filtering at the endpoint itself.

The basic building blocks of cyberinfrastructure itself is another area that poses critical research challenges. We have been facing the limitations of complementary metal–oxide–semiconductors (CMOS) and Moore's Law for some time. Additional research is needed to explore experimental and low power architectures, too. Examples of this include field-programmable gate arrays (FPGAs), neuromorphic chips, and quantum computing.

CHALLENGE 3: Modeling, simulation, and manufacturing across multiple scales

Modeling and simulation are central to our scientific understanding of the world and our ability to develop control and optimization technologies. Frequently, science and engineering applications are scaled down to fit the infrastructure, reducing scientific discovery and innovation. Solving this multi-scale problem is one of the grand challenges for cyberinfrastructure.

This has been an ongoing challenge in our understanding of dynamical processes – where varying magnitudes of length and time scales must be represented in the same problem -- for example in unsteady reacting flow simulations. Such simulations involve a complex interplay of hydrodynamics, turbulence and reaction chemistry. Applications of this behavior range from combustion and earth science models to drug delivery simulations. They all require high accuracy.

Modern materials science depends on advanced cyberinfrastructure to do predictive studies as well as molecular design and simulated testing. The customized toolsets are data-driven and highly dependent on multiscale computations to design the next generation of smart, strong materials.

Our deepest understanding of the universe is driven by advanced astronomical observations and the ability to run simulations that capture the entire cosmological volume. There is further need for the development of custom-built simulations to study the interaction of very-high-energy particles with matter.

CHALLENGE 4: Rapidly advancing biological and molecular sciences to improve human health

In the course of a single lifetime, we have gone from the discovery of DNA structure to the ability to copy and paste genes. There have been astounding advances in life and health sciences, and the progress is heavily dependent on cyberinfrastructure, since biological science and engineering is increasingly computational and interdisciplinary.

A fundamental research challenge is to uncover the relationship between molecular function and phenotype. We are able to sequence genomes for most organisms, but without a more complete understanding of proteins and their connection to phenotypes, we are not able to fully leverage the power of bioinformatics.

There are clear health applications to this work. A more complete understanding of the microbial genome along with advanced molecular simulations could pave the way to better drug design and even take a step toward fulfilling the promise of personalized and precision medicine.

Modeling and simulation work at the level of molecular dynamics and quantum chemistry requires significant computational resources, but can yield important insights into organic materials which may improve solar cells and advance our understanding of the origins of life.

Question 2 Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

A shift in the approach to national cyberinfrastructure is necessary to accomplish the challenges identified above.

1. Building a multi-purpose research cyber-infrastructure

Future breakthroughs are reliant on continued investment of national level resources in the path to exascale systems. That said, there are real limitations in an approach that primarily relies on "big iron" systems. Among Georgia Tech researchers who currently use large computational resources, there is general concern that they are highly oversubscribed, leading to long queues and wait times. More broadly, the perception is a general lack of resources to accommodate large simulations due to smaller jobs that require high-throughput computing. This problem is not likely to be addressed by reaching exascale capacity as there is essentially unbounded demand yet natural boundaries to scalability at many levels. Few researchers have access to funding to port code to new architecture introduced by these “big iron” systems. The national scale resources are also not well suited for small to medium-sized jobs and local institutional support is uneven

and inconsistent.

Our existing cyberinfrastructure is also limiting for researchers who need more data-centric systems. Many modern computational tasks are "embarrassingly parallel" and have strong scalability, but available computer clusters and HPC systems are not designed or optimized for such HTC workloads. Examples include data analytics and deep learning workloads. We must develop new systems that can more efficiently support data intensive applications. There are promising technologies for this including modern memory hierarchies, GPUs, and other heterogeneous environments.

2. Public-private partnerships

Public-private partnerships are necessary to increase the pool of resources dedicated to the execution of embarrassingly parallel workloads. Efforts like the open science grid are a good start, but they still rely on limited institutional resources and clear incentives to participate are required. A concerted effort is needed to migrate as much workload as possible to environments that provide hyperscale computing. This would include simplifying and rationalizing the ability for researchers to more easily use public cloud providers such as Amazon's AWS, Microsoft's Azure, and Google's cloud. When evaluating cloud resources for scientific computing, the full cost of local versus remote operations should be considered, along with cycles and hardware. Hyperscalar service providers have solid experience and motivation to operate highly resilient, geographically distributed, and low cost storage repositories and computational servers. If there were resource credits available to researchers which could be used for commercial cloud services as well as national, regional, and even local scientific computing, then we would have a better understanding of demand and efficiency needs.

3. Evolution toward greater agility, scale and resilience

A significant investment in software engineering is necessary to enable researchers to leverage the promise offered by public-private, multi-cloud based cyberinfrastructure. New technologies such as strong, secure container technology (or a similar virtualization platform) are a first step. However, researchers need a corresponding software infrastructure built on modern software engineering principles so more time is spent on asking and answering important questions, and less time on developing one-time technologies needed to maintain old workflows in new environments. Leveraging emerging high level services (such Function as a Service) may help with scaling computations across multiple clouds.

We note that legacy codes generally do not map well to the cloud and new architectures. We need to support algorithmic development which is less dependent on low latency systems. In an ideal scenario, the system, not the code, would handle platform translation, optimization, and automatic scaling. This is potentially something that machine learning can provide. Sustainable code is the only path toward a more universal portability model that can allow algorithms to be built once and then run on many platforms.

Pay-as-you-go computing and storage is a very different budget mindset. Researchers must adapt to maintain control of their usage and prevent errors from eating up their allocations. Curating datasets can be an unsustainable proposition in such an environment without a strong subsidy model or other overage protections.

4. The infinite data problem

Low cost, long term, reliable large scale data storage is a recurring demand across researchers. A sustainable solution could be organized around multi-petabyte storage facilities operated by approved public-private consortiums, following common guidelines. Such a solution would contribute toward addressing problems related to data integrity, portability, and reproducibility. Designation of authoritative repositories for specific datasets also helps reduce the risks that result from data duplication.

Another limitation is the location of the data itself. The biggest obstacles remain cost and time associated with moving data over a wire. As described above, many datasets are generated in a distributed way (e.g. sensor networks, genomes). Current architectures typically require the capture and migration of this data in order to do analysis. This impedes the use of real-time analytics which could be accomplished at the data sources with a more coordinated yet distributed infrastructure. Beyond more compute and storage, this implies further development and investment into new technologies (hardware and software) such as inline data analysis (e.g. specialized machine learning and/or artificial intelligence frameworks) to detect important data features, and communication avoidance algorithms.

There is an ever-growing challenge associated with the sharing of large-scale datasets, ranging up to petabytes in size. We must make continued strategic investments to ensure sufficient network bandwidth and speed. Advanced networking technology like software defined networking is also an important part of the solution.

5. Cybersecurity at the core of research cyberinfrastructure

Sharing and disseminating datasets while ensuring security and compliance is a major concern among researchers. Multi-institution and multidisciplinary research collaborations tend to be quickly hampered by security and compliance requirements.

One possible solution is the creation of approved secured facilities or other data clearinghouses, capable of supporting access by distributed teams, and various modes of computing (including graphics accelerated design and visualization). Given that data is increasingly expensive to move, we must find ways to bring the computing resources to the data anyway. It may also make sense to leverage machine learning techniques to develop automated processes capable of de-identifying and/or certifying "clean" datasets for faster, more consistent access by researchers.

Continuous and sustained investment in security is a must, but without usability we will not be able to advance science at all. Verifiable trust models (e.g. emerging technologies with a custody chain auditing and cryptographically secured ledgers) may be a path forward.

6. Emerging Research Cyberinfrastructure

A final limitation of existing cyberinfrastructure is that there is a barrier to testing radical new architectures “beyond Moore’s Law” (e.g. massively many-core, memory-centric compute, neuromorphic, quantum computing, FPGA, open-source hardware, moving the compute to the data) at scale. A carefully developed proving ground which would allow for experimentation and tolerate a higher risk of failure would be a welcome addition to the national conversation and could provide an on-ramp to encourage rapid adoption of promising disruptive technologies.

A public-private coordinated approach to such investment will also ensure that researchers can leverage future massively distributed, low power, and edge-computing ecosystems.

While most researcher demands are for resources in support of large parallel, parametric or data-centric computation, a new category of needs is emerging to support further cyber-physical related research. Investigation involving robot-human interaction calls for dedicated facilities supporting secure tele-operation, augmented reality environments, and high-fidelity real-time exchange of large amounts of data.

Question 3 Other considerations (maximum ~1200 words, optional): Any other relevant aspects, such as organization, process, learning and workforce development, access, and sustainability, that need to be addressed; or any other issues that NSF should consider.

1. Workforce development

One of the greatest risks to our current cyberinfrastructure is that it will become virtually unusable due to a lack of researchers and staff with an appropriate background in both the technology and the science. Many researchers placed a strong emphasis on the importance of investing in the human capital who operates research cyberinfrastructure, and also investment in various forms of support to researchers.

A major challenge is that students who have the skill and interest to do scientific computing are difficult to attract into such jobs, including national laboratories and smaller research groups. Part of the difficulty lies in the marketplace. We are competing with an industry which can offer significantly higher compensation, and at times, even competing within the science and research community as skilled research scientists move between institutions.

There is another obstacle. Successful candidates need to do more than just code – they must have depth in one (or more!) domains of research. The next generation of cyberinfrastructure will require a steady pipeline of these "unicorns." Many of these career paths are suitable for graduate students to pursue, so perhaps we need to find a way to reward faculty members for producing such students.

In order for the future cyberinfrastructure to be robust and provide value to the scientific community, we also need people with practical experience in working with large data sets. We need software engineers who have strong development skills to implement new algorithms and scalable codes that leverage emerging architectures. And of course, researchers need to have a solid understanding of cybersecurity and policy compliance as applied to unique research environments.

The training needed for this workforce will change rapidly and continued professional re-tooling will become the norm. Such rapid curriculum change could happen via partnerships that allow universities to easily modify courses to better reflect future technology directions. Another approach to the workforce challenge is to build small, local or regional teams of developers, technologists, data scientists, researchers, policy experts, etc. who are available to assist on multiple projects.

Finally, we must continue to develop and integrate fundamental concepts needed to build or use new radical computing technology such as quantum computers.

2. Scientific cyberinfrastructure as a public good

One of the best ways to address the workforce demands is to grow the pipeline of students entering science and engineering careers. Population trends indicate that a good approach is to increase the diversity of underrepresented groups. This will have the additional benefit of expanding the diversity of thinking about cyberinfrastructure, leading to improvements in technology, process, and results.

Cyberinfrastructure doesn't reach all science and engineering students and researchers equally. It would be potentially transformative to establish a pool of resources that could be accessed by K-12 students, and by teachers and researchers at colleges and universities who are not traditional users of advanced computing. Some science gateways are already in place which take a good step in this direction, but a more deliberate and broad approach is strategic.

Outreach to citizen researchers, especially in the context of edge computing and sensor driven research, is another way to address the issue of access. We can enable private citizens to more easily access data and contribute data by tying sensors into large sensor fabrics. Science is a human endeavor and cyberinfrastructure is another opportunity to engage emerging scientists everywhere.

Asian countries (especially China) are steadily increasing investments in high performance systems, and there is a risk that the United States may appear to have a reduced leadership position should we redirect resources away from traditional HPC to more data centric models. However, we would be better served as leaders in science by balanced strategic investments than we would be by maintaining spots on top ten lists.

3. Data management

Scientific and research data management is another limitation of the current environment. There is generally a lack of academic-private accepted standards and best practices around data format, data storage, best practices, metadata (which is as important as the data itself), sharing frameworks, and data governance.

The NSF should consider playing a greater role in guiding the development and promotion of an all-encompassing data life cycle approach which could be achieved through a strong public-private partnership model. Under such an approach, a national Science and Engineering data consortium could establish universally-recognized standards and best practices. There are independent bodies (IEEE, IETF) which have shown that successful standards can create efficiencies in technologies.

Another aspect to data management are the real costs which are not well-accounted for. There must be tighter integration between compliance and research in order to break down barriers that prevent data from being collected or mined. An understanding of the compliance issues by the researchers is a necessary training; however, those dealing with compliance also need a better understanding of the workflows needed to perform research, especially in data-driven science. This encourages collaboration when building environments that are productive to novel research.

4. Sustainable funding models

We must protect our investment in capability hardware, but also begin to move toward data-centric computing. We must also maintain a small but significant portion of the cyberinfrastructure portfolio for experimental architectures. The challenge is how to pay for it.

There are national resources (e.g. XSEDE) which are generally funded with large federal investments. States and cities should be engaged since they may be interested in partnering on the funding of regional resources (e.g. Big Data Hubs) if they can be shown to have economic impact. Local institutional resources are another potential source of funds.

We mentioned public-private partnerships (e.g. academic and industry, government and corporate) earlier, and that is a key component.

Everyone benefits from investments in cyberinfrastructure, so it makes sense that we should investigate mechanisms for broader support.

Finally, we point out that the costs of fabrication are prohibitive and the slowdown of Moore's Law means that the lifecycle of products in the near future will be three to eight years rather than two to three years. Economic growth will be based on innovative services on hardware platforms that are reliable and secure. Services will be based more on creative use of data rather than faster hardware. This trend reinforces the need to diversify the cyberinfrastructure portfolio in the United States, and the NSF should lead the way.

5. Acknowledgements

As mentioned above, this RFI response was distilled from the enthusiastic responses of more than 50 faculty and staff at Georgia Institute of Technology. We appreciate the input of:

Srinivas Aluru, David Bader, Peter Brecke, Jean-Luc Bredas, Rob Butera, Thomas Conte, Didier Contis, Steve Cross, Rich DeMillo, Chaitanya Deo, Lizanne DeStefano, Magnus Egerstedt, C. Ross Ethier, Dan Forsyth, Martha Grover, JC Gumbart, Felix Hermann, Troy Hilley, Yongtao Hu, Bartosz Ilkowski, Jeff Jenkins, Surya Kalidindi, Mark Keever, Kostas Konstantinidis, Julia Kubanek, Satish Kumar, Rachel Kuske, Joseph Lachance, Pablo Laguna, Lew Lefton, David Leonard, Tim Lieuwen, Hang Lu, Jimmy Lummis, Paul Manno, David McDowell, Suresh Menon, Perry Minyard, Nepomuk Otte, Sebastian Pokutta, Jason Poovey, Dana Randall, Justin Romberg, Chris Rozell, David Sherrill, Deirdre Shoemaker, David Sholl, Jeffrey Skolnick, Madhavan Swaminathan, Julie Swann, Ignacio Taboada, Rich Vuduc, Yan Wang, Joshua Weitz, John Wise, Chris Wright, Sudhakar Yalamanchili, and Minami Yoda.

Consent Statement

- "I hereby agree to give the National Science Foundation (NSF) the right to use this information for the purposes stated above and to display it on a publically available website, consistent with the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode)."